Policy Hyperparameter Exploration for Behavioral Learning of Smartphone Robots(Abstract_要旨)

AUTHOR(S):

Wang, Jiexin

（ 続紙 1 ）

京都大学	博士（　　情報　　学）	氏名	王　潔心
論文題目	Policy Hyperparameter Exploration for Behavioral Learning of Smartphone Robots（スマートフォンロボットの行動学習のための方策ハイパーパラメータ探索法）		

（論文内容の要旨）

Robots constitute the counterpart of humans. Scientists in the field of robotics and artificial intelligence have spent decades on inventing human-like intelligence and enabling robots to perform biological agent-like behaviors in a wide spectrum of applications from assisting humans to understanding animal's evolution. Our main research goal is to understand emergence of behaviors and minds of biological creatures, by realizing a wide variety of robot's behaviors. Such issues include investigating how an individual learns, how a group evolves, and how a reward system is developed therein. To do so, we developed a smartphone-based robotic platform and a novel reinforcement learning algorithm for the smartphone robot to perform various behaviors.

Current robotic research platforms, including humanoids, iCub and NAO, and small desktop miniatures such as Khepera and e-puck, are either too expensive, hard to maintain, or equipped with limited computational power and sensors. To overcome these limitations and enable variety of behaviors with more degree of freedom, we developed an affordable high-performance robotic platform, namely, a smartphone balancer as an individual. The balancer's behaviors include basic mobility: standing-up, balancing and stable running, and integrated complex behaviors: foraging, collision avoidance and communication. This platform provides a practical testbed for designing optimal control and reinforcement learning algorithms, has high applicability to behavior-based robotics and multi-agent research, and also benefits education and hobby use.

In this thesis, we first proposed the design and control architecture of the smartphone balancer and test their feasibility of allowing the balancer to perform standing-up and balancing behaviors under the hardware constrains. Generally, it is difficult for small-sized robots to mount motors emitting high torque. We proposed a spring-attached and wheeled inverted pendulum model of the smartphone balancer to overcome the difficulty of the motor torque limitation. We scanned different combination of motor torque, spring coefficient and frequency of periodic controls in both simulator and actual hardware system. Results showed that the attachment of the elastic bumper simplified the control and assisted the robot to achieve the required behaviors successfully.

For a single agent to acquire basic behaviors efficiently in practice, we developed a deterministic policy search method, EM-based Policy Hyperparameter Exploration (EPHE). EPHE integrates policy gradient method (PGPE) and EM-based policy search framework to avoid gradient calculation and learning rate tuning. It assumes a prior distribution over the policy parameters and updates the hyperparameters in a closed form solution of maximizing the lower bound of expected return. This results in an updating rule to weigh received returns, when the prior distribution is from exponential

family. However, updating with the fully observed dataset decelerates the learning performance, due to disturbance from non-preferable samples.

There are various discarding rules and weighting schemes based on the return history to preserve informative samples. We first implemented a K-elite selection heuristics onto EPHE (EPHE-K) to discard non-preferable samples under a fixed baseline, because similar heuristics were adopted in previous policy search methods like PoWER, FEM and CEM. Methods rescaled the returns by means of a convex transformation, such as CMAES and REPS weighting schemes, can also be equipped to our learning framework; CMAES reweighted the samples according to a logarithm transformation, and REPS used an exponential transformation. However, all these methods have to determine the fixed baseline pre-requisitely. To avoid tuning the K parameter of the fixed baseline and hence set the baseline adaptively, then we introduced an adaptive baseline scheme on EPHE (EPHE-AB) to discard non-preferable samples below the average of the return history. We also examined several baseline settings including the mean and one or two standard deviation(s) from the mean.

Next, we implemented EPHE-K and EPHE-AB with the baseline settings of the mean, and one or two standard deviation(s) from the mean in three simulation tasks; they are pendulum swinging-up with limited torque, cart-pole balancing, and standing-up and balancing of our smartphone balancer. Results showed that EPHE-K outperformed the previous policy gradient methods like PGPE and Finite Difference. And EPHE-AB outperformed EPHE-K, EPHE with CMAES weighting scheme, EPHE with REPS weighting scheme, PGPE, NES, and FEM especially in the early stage of learning. It was shown that the setting the adaptive baseline at the mean was more effective in focusing on preferable samples than others, leading to steady decrease in the number of discarded samples during learning. We further implemented EPHE-AB with the mean baseline on the real smartphone balancer system, for realizing its standing-up, balancing, and visual target approaching behaviors. Results showed that our EPHE-AB outperformed PGPE, so to allow the smartphone balancer to successfully achieve the behaviors.

Finally, we implemented the applicability of our EPHE-AB to acquisition of high-level behaviors by the balancer. Results showed our smartphone robot achieved foraging and communication behaviors event under its low-cost hardware settings.

注)論文内容の要旨と論文審査の結果の要旨は１頁を３８字×３６行で作成し、合わせて、３，０００字を標準とすること。
論文内容の要旨を英語で記入する場合は、４００～１，１００wordsで作成し
審査結果の要旨は日本語５００～２，０００字程度で作成すること。

（続紙 ２ ）

（論文審査の結果の要旨）

強化学習は、試行錯誤を通して人工システムの制御則を獲得するための枠組みであり、近年、ビデオゲームや囲碁において人間よりもうまく振る舞うことのできるシステムに用いられるなど注目を集めているが、ロボットへの応用には未だ問題が残っている。一つには、適度な複雑さを持った実験プラットフォームがなく、人工知能分野の研究者がアルゴリズムを検証することが容易ではないことである。またロボットへの応用では、少ない試行数からの制御則学習が必須であるのみならず、確率的な制御則を使った学習試行は安定性の観点で望ましくない。こうしたロボットへの応用における重要な点は、ゲーム学習で有効性が示された強化学習法アルゴリズムでは、あまり重視されてこなかった。本研究では、様々な分野の研究者にとって使いやすいロボット実験プラットフォームを開発すると共に、実ロボット上で実装可能な、効率の良い強化学習アルゴリズムを提案している。これまでに、以下のような成果を得ている。

（１） 決定論的な制御則を用いた学習が可能な、繰り返し統計学習（EM）アルゴリズムに基づいた強化学習法を提案した。決定論的な制御則を使うことで、従来の確率的な制御則に起因する試行系列のばらつきを抑えることができる。またEMアルゴリズムを用いることで学習率などのメタパラメータを調節する手間を軽減した。さらに、効率の良い制御学習の実現のため、学習試行を選択するための閾値を自動調節する手法を開発した。これにより従来のEMアルゴリズムを用いた強化学習法よりも少ないデータ量からでもタスクを達成する制御則を獲得できることを、いくつかの標準的なベンチマーク課題を通じて確認した。

（２） スマートフォンに車輪をつけた倒立二輪構造の移動ロボット（スマートフォンロボット）を開発した。このロボットは、これまで人工知能研究者が使用してきた小型の移動ロボット実験プラットフォームと比較して、大幅に安価に製作できるだけでなく、倒立を維持しつつ空間移動するなど、ダイナミックな制御が必要となる点で適度に複雑である。また、スマートフォンの無線充電機能を使って充電可能であり、エネルギー補充をしながら直立移動を行う、などの特徴を有し、生物模倣型ロボットのテストベッドとしても利用可能である。

（３） 倒立二輪型のスマートフォンロボットに必要な基本行動である倒立安定化、カメラ画像に基づくターゲット位置への到達、充電ステーションへの移動および充電、二台のロボット間のコミュニケーションなどの行動を（１）で開発した強化学習法により実現した。これらの行動はシミュレータを利用せず実ロボットを直接動かすことによる、比較的少ない試行によるデータからでも学習されており、（１）のアルゴリズムの効率の良さを示している。また、これらの行動の実現は、

今後、生物模倣型ロボットによるコミュニケーション創発や、小型群ロボットの集団進化など人工知能研究への展開が期待される。

以上を要するに、本論文は人工知能など様々な分野の研究者にとって利用しやすいロボットプラットフォームを開発した点、および、実ロボットから得られる、少ない試行数から適切な制御則を学習できる、効率の良い学習制御のアルゴリズムを提案し、評価した点で重要であり、博士（情報学）の学位に値するものと認める。

平成29年2月27日に論文内容とそれに関連した口頭試問を行った結果、合格と認めた。

注）論文審査の結果の要旨の結句には、学位論文の審査についての認定を明記すること。更に、試問の結果の要旨（例えば「平成　年　月　日論文内容とそれに関連した口頭試問を行った結果合格と認めた。」）を付け加えること。

Webでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。
要旨公開可能日：　　年　　月　　日以降